DOCUMENT RESUME

ED 299 323 TM 012 358

AUTHOR Coffman, William E.
TITLE Research in Writing Assessment.
PUB DATE 16 Apr 88
NOTE 9p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Information Analyses (070) -- Viewpoints (120)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Educational Research; Elementary Secondary Education; Essay Tests; Evaluative Thinking; Higher Education; Latent Trait Theory; *Measurement Techniques; *Research Reports; Test Reliability; Test Validity; Writing (Composition); *Writing Evaluation; Writing Skills
IDENTIFIERS Criticism; National Assessment of Educational Progress

ABSTRACT

Five papers on the assessment of writing ability presented at the Division D meeting are critiqued. It is suggested that it is not possible to conclude--as has Larson in his study of grades 4, 6, and 8--that National Assessment of Educational Progress results do not give evidence of growth in writing ability. Dovell and Buhr, examining approaches to a system for generalizing essay writing scores in a state testing program, should perhaps have asked how much confidence one can have in decisions based on such scores. The examination by Cantor and Hoover of the reliability and validity of essay tests of writing ability was not designed to be generalizable. Swartz, Patience, and Whitney have asked if essays can contribute unique variance to the measurement of writing ability, but the data do not support their conclusions. The examination of item response theory by Ackerman, despite flaws in the analysis, does show it as a powerful tool. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED299323

RESEARCH IN WRITING ASSESSMENT

Critique, Division D, AERA
Wednesday, April 16, 1988 San Francisco

Discussion: William E. Coffman
University of Iowa

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

William E. Coffman

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM012358

RESEARCH IN WRITING ASSESSMENT

(Critique, Division, D, AERA. Wednesday, April 16, 1986. San Francisco)

Discussion: William E. Coffman
University of Iowa

After reading each of these five papers, I recalled a statement in the Cantor and Hoover paper in which they summarized their review of the literature. They wrote:

> Twenty or more correlational and experimental studies published since 1966 specifically address the reliability and validity of the essay score. The purposes, and hence the designs, of these investigations vary widely, as do their results. Sample sizes range from 5 to over 10,000. Interrater reliabilities range from .366 to .92. Estimates of score reliability .....range from .221 to .81. Correlations between essay scores and scores on objective measures of language skills range from .20 to .81.

The reviewers do not explain why these vast differences in findings, and they don't explain how their own findings might help to resolve the differences. It's puzzling, therefore, to find them concluding, in the next paragraph, that "we are now able to develop direct measures of writing with a fair degree of confidence in their reliability." One wonders whether they are summarizing their review of the previous literature or whether they are stating a conclusion based on their own study. If it is the latter, do the other papers in this session tend to confirm or to contradict their conclusion?

Larson asks about the appropriateness of NAEP exercises for assessing writing ability in grades 4, 6, and 8 in Montgomery County, Maryland and concludes that they aren't appropriate. One may question, however, whether the methods of analysis really provide answers to the questions he asks. He reasons that, since the correlations between scores obtained on essays administered in the fall with those administered in the following spring to the same students are low, the essay test scores are so unreliable as to be useless for individual measurement. But are correlations between fall and spring measures valid indicators of measurement reliability? I would argue, rather, that the purpose of teaching is to invalidate measures obtained prior to teaching. If the teaching is designed to remove deficiencies, pretest-posttest scores might very well be low. On the other hand, even if we accept the correlations reported by Larson as reasonable estimates

of the score reliability, such scores might actually be providing useful information.

I have argued elsewhere (Coffman, 1966) that, if a particular essay product can be considered a random sample from the population of such products the student might be asked to produce, then the validity of the scores can be estimated by the square root of the reliability coefficient. Validity coefficients of .539 ($\sqrt{.29}$), .592($\sqrt{.35}$), and .624 ($\sqrt{.39}$) estimated from Larson's data compare favorably with validity coefficients obtained from objective measures of writing ability.

Larson also concludes that the correlations between essay scores and scores on objective tests for the Montgomery County students are so high that there is no evidence the essay scores are measuring anything not already measured by the objective tests. Unfortunately, he attempts to interpret the raw correlations. The measures that should be compared are correlations corrected for attenuation, and Larson doesn't provide the reliability coefficients necessary to make such corrections.

Let's assume, however, that the essay scores from the NAEP prompts are comparable in reliability to those obtained by Godshalk, Swineford, and Coffman (1966) and examined in detail by Coffman (1966). I found that the reliability of the sum of three essays each read by two different raters would be .647. The pre-post correlation of .58 reported by Larson, when corrected for attenuation using this estimate of reliability, would become .896. In contrast, assuming a reliability of .92 for CAT reading and .95 for CAT language (not unreasonable assumptions), the corresponding corrected correlations with the essay pretest scores would be only .726 and .765 respectively. In contrast, the corrected correlations between CAT reading and CAT language becomes .888. It might very well be that the NAEP prompts are providing Larson with some useful data.

Regarding the conclusion that the essay scores do not provide evidence of growth in writing ability, it is impossible from the data provided to conclude this. An alternative interpretation is that there really has been no growth from fall to spring in a single year. It would be helpful if the objective measures of writing ability had also been repeated in the spring.

Dovell and Buhr examine five approaches to the development and analysis of a system for generating essay writing scores in the context of a state testing program. I think it is probably going too far to dignify these

approaches by the term models. It might be better if they had simply discussed the issue of how one can determine how much confidence one can have in decisions based on test scores collected in the several ways. I have emphasized the word decisions because when the purpose of a test, as in this program, is to place individuals into two classes, pass or fail, the critical question is this: "What is the probability that an individual that should pass will be classified as fail or that one that should fail will be classified as pass?" This is a question in decision theory, and I see no reference to decision theory in the paper.

The paper does recognize that it would be good if one could have conducted a generalizability study and obtained estimates of the various sources of error that one might expect to be operating. Given such information, one could make more informed decisions regarding the way to organize the program. One could also use such estimates to generate the probability estimates needed to evaluate the decisions.

Actually, what the paper describes are two essentially aribtrary decision rules, one that fails any paper that does not receive a total score of 4 when papers are read twice on a four-point scale, the other that establishes a pass-fail point for an objective test of writing ability and bases pass-fail decisions on a combination of essay and objective scores. The authors report that when all borderline or widely discrepant essay scores are refereed, and when essay scores are expanded to a scale that seems to be much more accurate than it really is, it doesn't make much difference whether you combine essay and objective test scores or whether or not you adjust the essay test scores on the basis of differences in objective test scores from administration to administration. In other words, if we avoid the issue of misclassification by fiddling with the numbers, we can make ourselves more comfortable. We really ought not to be doing this. What we ought to be doing is asking the question with which I introduced this section of my comments: "What is the probability that a student will be misclassified?"

Cantor and Hoover examine the reliability and validity of essay tests of writing ability by collecting and analyzing a large amount of data. Unfortunately, they did not design their study as a generalizability study, so it is difficult to be sure how far one can apply their findings to the typical situation in which the tests will be used. What we do know is that if we collect essays using these particular test prompts (or others like

them) and have them rated by the pool of professional raters that scored the papers in this study (or others like them with similar training), using the scoring directions provided by the Iowa Testing Programs, the scoring reliability will be very high, and also that one can provide a reasonably accurate estimate of the size of the typical error of measurement. We can't be sure what will happen if typical classroom teachers attempt to follow the scoring procedures.

One thing this study does confirm: the critical issue in essay testing is not rating reliability; rather it is score reliability -- or, more generally the question of what can the obtained scores be generalized to. It is possible, given well prepared directions and careful training, to obtain highly reliable ratings (.78 to .97; median .90 for single ratings). Since student performance from time to time and from prompt to prompt can be expected to vary, the reliability of a single essay, regardless of the reliability of the rating, will be relatively low (within mode -.47 to +.77 with a median of .465; between mode +.16 to +.48 with a median of .37). Cantor and Hoover probably make too much of the differences between the .465 median reliability within mode and the .370 median reliability between mode. In Session 43.06 on Saturday afternoon, Breland will be reporting that the variance error for topics is more critical than the variance error for mode. Note also that even with the elaborate steps taken to increase reliability of rating, the result across modes found by Cantor and Hoover (.370) is essentially the same as the estimate reported by Coffman in 1966 for the sum of two ratings for a single topic (.380).

Unlike Larson, Cantor and Hoover do correct correlations between essay and objective test scores for attenuation and conclude that their tests are contributing some unique variance to the measurement of writing ability. The raw correlations they report between essay and objective test scores, are somewhat lower than those reported by Larson. One hypothesis to account for this is that they did develop scoring directions that focused on those aspects of written communication that might not be measured by objective tests. It's not clear whether such a focus would be found in the directions for scoring the NAEP essays, but I doubt it. Certainly it does seem inefficient to try to measure with essay samples those aspects of writing such as spelling, punctuation and capitalization, etc. that are more efficiently measured by objective tests.

Swartz, Patience, and Whitney have investigated the use of essays at the high school graduation - college entrance level. They seem to have been concerned primarily with the question of rating reliability and with the issue of whether or not essays can contribute unique variance to the measurement of writing ability. Unfortunately, the conclusions they draw are not supported by the data they report.

They do report that it is possible to obtain high agreement between ratings obtained by raters working in different settings and with backgrounds of experience with papers at different ranges on the score scale. This finding tends to support the conclusions I draw from the data of the Cantor and Hoover study: the critical issue in essay testing is not reliability of ratings nor is it even score reliability. Rather, the critical issue is whether or not essay scores can contribute unique valid variance to the measurement. Swartz, Patience, and Whitney conclude that they can. However, they base their conclusions on raw correlations, not on correlations corrected for attenuation, as did Cantor and Hoover. Actually, without more information, especially the reliability coefficients for the tests involved for the samples involved, it is impossible to evaluate their claim. However, using the estimate of .60 for the reliabiity of the essay score and a reliability of .92 for the objective writing score, one obtains a corrected correlation of .939 between the essay scores and the objective scores for 1983 sample. This does not leave very much in the way of unique variance for the essay examination.

One other point needs to be made. Swartz et al. seem not to recognize that if they had computed the intraclass coefficient rather than the product moment coefficient of reliability, the estimate would have included any mean differences among ratings as part of the error of measurement. The intraclass coefficient is always preferable to the product moment coefficient when reporting reliability coefficients unless you are generating generalizability coefficients from analysis-of-variance estimates of error components.

The final paper by Ackerman demonstrates how the IRT model permits a much more detailed examination of the contribution of individual items or individual ratings to a test score than does the classical model if the IRT model fits the data. However, in interpreting the kinds of estimates of information presented by Ackerman, one needs to recognize that the IRT model, like the classical model, focuses on reliability, not on validity. When Ackerman writes that the essay score for spelling provides more information than the objective score, one needs to remember that there is

an underlying assumption, i.e. that these two spelling tests are measuring an underlying spelling ability that is validly represented by these particular samples. Let's suppose that certain pupils in the sample have learned how to avoid using in an essay any words they do not know how to spell and that other students have not. In that case, how much confidence can we have in the sample of writing as a sample of spelling ability? Notice also that each rating of the single essay is treated as an item for purposes of analysis even though the same rater may be providing five different judgments based on the same sample of writing. It would have been much more informative if at least two samples of writing had been collected and if each judgment were made by a different rater.

Note also that in order to obtain essay scores for the same elements as those assessed by the objective test scores, Ackerman had to require a scoring procedure that focused on error counts rather than quality judgments. Even then the raters did not agree very well. By fitting the essay scoring into the Procrustean bed of the objective test, Ackerman eliminates the possibility of obtaining unique information from the essay. Notice also that the procedure of placing the two sets of scores on the same scale assumes that the combined sample of objective test data and essay score data serves to define the underlying ability. The IRT analysis provides data for testing whether or not the items are measuring a single dimension of ability; however, there is no evidence that Akerman applied such a test. We aren't told how well the data fit the model, but we are told that in many cases, the item scores do not provide very much information about whatever ability is being measured by the collections of items.

In conclusion, the IRT model appears to be a powerful tool for analyzing test data. In many cases, it will tell you more about what's wrong with the test than what's right with it. And if the data we've collected look to us as if they ought to be measuring what we intended to measure but they don't seem to be providing much information, we're going to have to do a lot of hard thinking about the reasons why. The IRT model just won't do our thinking for us.

REFERENCES

Coffman, W. E. (1966). On the validity of essay tests of achievement. Journal of Educational Measurement, 3, 151-156.

Godshalk, F. I., Swineford, F., & Coffman, W. E. (1966). The Measurement of Writing Ability. New York: College Entrance Examination Board.